



Data Science Studio初步使用报告

1. Data Science Studio (DSS)

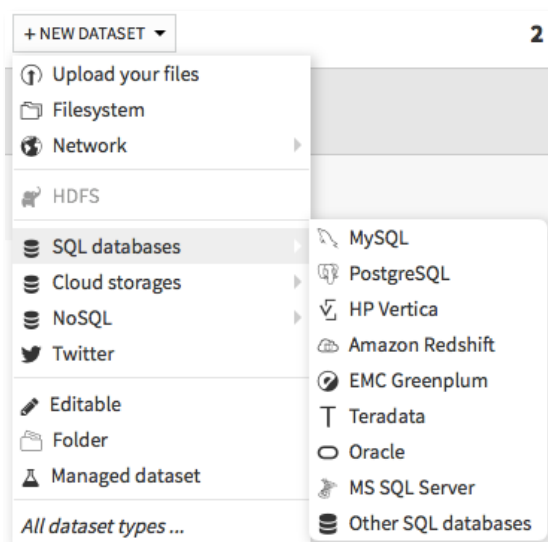
Data Science Studio是Data iku公司研发的一款数据分析软件。Data Science Studio整合了从数据整合，数据清洗，可视化分析，机器学习以及到最后的发布流程。它把数据库，可视化分析，机器学习等大数据中使用到的工具都模块化了，提供了一个统一的操作界面。

2. Flow

在Data Science Studio中，所有的操作步骤都可以被记录下来，作为一套固定的流水线，类似于写程序。对于新数据，我们就可以快速使用原有的流水线进行处理。首先我们需要介绍一些概念：Dataset, Recipe, Machine Learning。

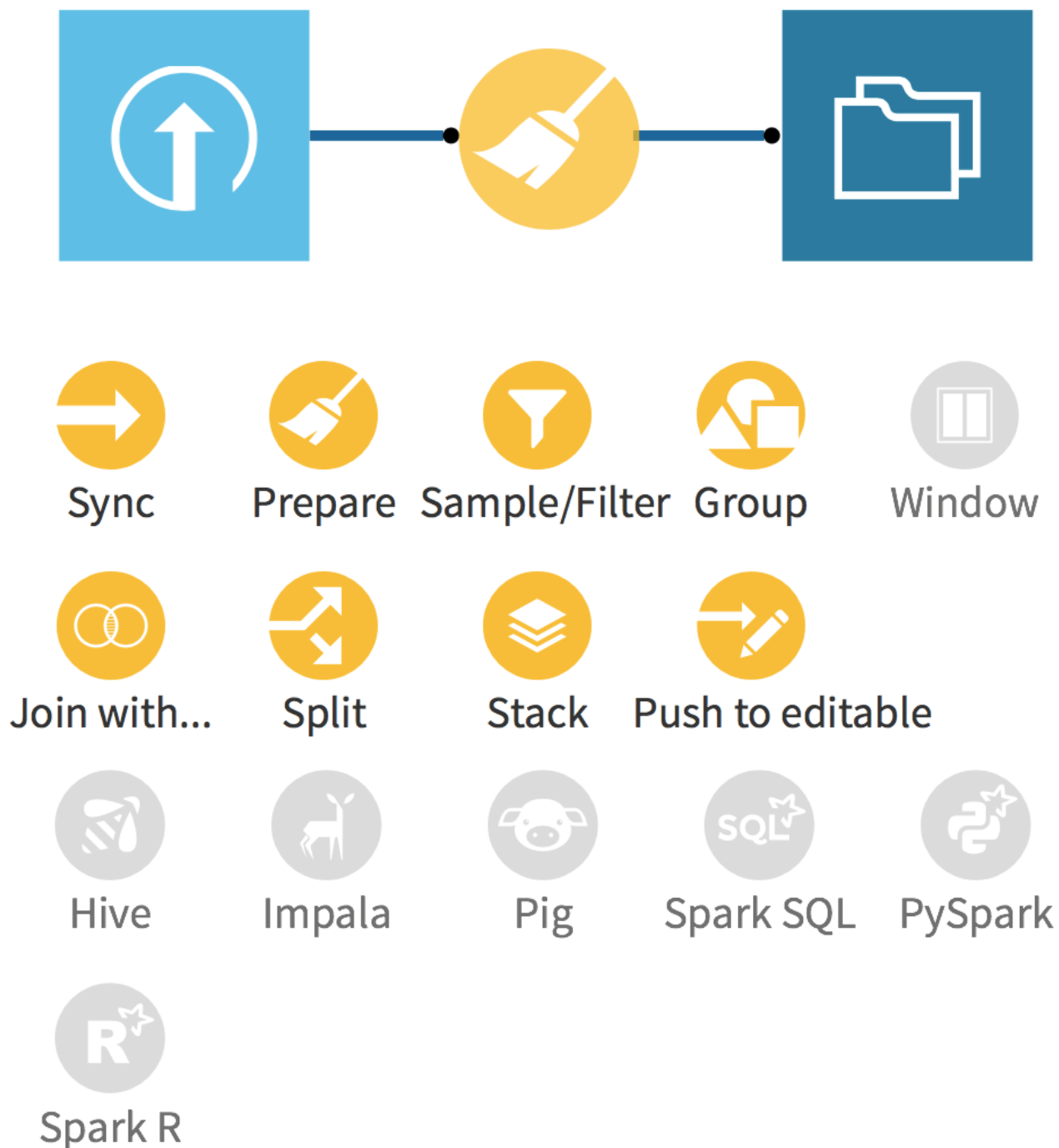
2.1. Dataset

使用Data Science Studio的第一步就是提供数据，Data Science Studio支持一系列数据库连接和离线文件，数据库包括MySQL, PostgreSQL, Oracle, Microsoft SQL Server, MongoDB和Twitter (Streaming API)等，离线文件包括基于文件系统, HDFS, Amazon S3, HTTP, FTP, SSH等协议下的文件，文件格式包括CSV, EXCEL, JSON, MySQL Dump和Apache Combined log format。一个dataset的图标是一个蓝色的方块，其中的图标表明了他的来源。左边所示的是一个来自上传文件的数据集，右边是一个上传按钮，显示了可以连接的各种数据源及图标。



2.2. Recipe

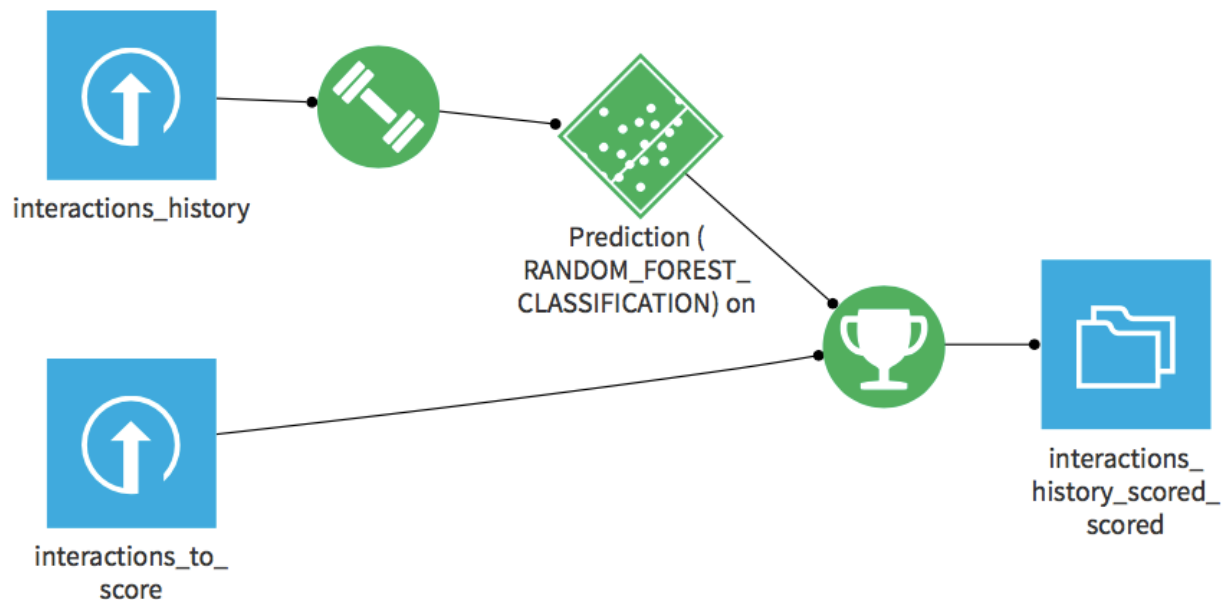
Recipe是对数据的一次处理，每对数据做一次transformation, aggregation, join等都会产生一个recipe。所以一个数据集经过一个recipe就生成一个新的数据集，可以说前者是recipe的输入，后者是recipe的输出。一个recipe由Data Science Studio中的工具可视化实现（比如一些数据清洗），或者是一些编程语句（）。下图所示的就是在从一个数据集经过Data Science Studio的工具得到的新数据集的过程，其中recipe展示为黄色的圆形（通过Data Science Studio实现的处理），不同的图案代表不同的含义。如果是通过编程语句实现的，就会显示为红色圆形，图案是相应的编程语言（由于没有连接相应的数据库，截图中的图案都是灰色的）。



2.3. Machine Learning

Data Science Studio系统中也集成了一些机器学习的工具，可以直接用来预测某些值。我们可以不断调节参数来提高模型的指标。下面是两个模型的截图。机器学习的过程使用绿色显示，杠铃代表模型训练事件，散点图代表模型计分，奖杯代表把模型应用到新的数据库中。

<input type="checkbox"/>	Random Forest with 84 trees 2016-04-27 - 14:37:44	<div>first_item</div> <div>country = United States</div> <div>age</div> <div>country = China</div> <div>campaign = True</div> <div>campaign = False</div>	<div>ROC AUC: 0.961</div>	<div>Details</div> <div>Random Forest with 84 trees</div> <div>◆ Active version</div>
	Random Forest with 84 trees 2016-04-27 - 14:29:31	<div>first_item</div> <div>country = United States</div> <div>age</div> <div>country = China</div> <div>campaign = True</div> <div>campaign = False</div>	<div>ROC AUC: 0.961</div>	<div>Details</div> <div>Random Forest with 84 trees</div> <div>Trained in 5 seconds on 23489 records</div> <div>MAKE ACTIVE</div>



2.4 Flow

Dataset, Recipe和Machine Learning构成了整个Flow（在网站中只说到了Dataset和Recipe），这构成了我们数据处理的整个流程，通过Flow，我们可以快速了解数据流水线式如何构成的。下面两张图是一些Flow的例子。

3. 其他一些功能

其他一些功能包括数据清洗，图标制作，内容发布。

3.1 数据清洗

在2.2 Recipe中提到的，我们可以通过Data Science Studio完成一些数据清洗。下面简单介绍一下。

3.1.1 导入数据

New UploadedFiles dataset

New dataset name

haiku_shirt_sales_1

CREATE

Connection

Preview

Schema

Advanced

user_id	departement	birth	order_id	order_date	total	nb_tshirts	tshirt_price	category
			Filled: 999 - Empty: 1					
a4dc1548af		11/9/1970	TSG-1200493	2013-12-07 15:27:00	71.7	3	23.9	White T-Shirt M
e8f6f7853c	91	1/31/1983	TSG-1200500	2014-01-22 08:28:00	19.9	1	19.9	Black T-Shirt M
5802162c20	46	6/5/1958	TSG-1200502	2014-03-22 21:07:00	19.9	1	19.9	Black T-Shirt F
67dc195843	92	12/30/1899	TSG-1200503	2014-09-07 09:03:00	16.7	1	16.7	Black T-Shirt M
7c52e0eab6	78	12/19/1981	TSG-1200505	2014-05-20 15:25:00	16.7	1	16.7	Black T-Shirt F
ef433d39c6	71	1/23/1991	TSG-1200508	2013-12-22 14:09:00	16.7	1	16.7	Black T-Shirt M

Type

Separated values (CSV, TSV, ...)

UPDATE PREVIEW

REDETECT

Quoting style

Excel style

How quoting and escaping is handled

Separator

\t

Can be a single character, or a Unicode escape sequence like \u0001

Quoting character

"

Character to enclose cells containing the separator. Can be a single character, or a Unicode escape sequence like \u0001

Skip first lines

0

☒ Parse next line as column headers

Skip next lines

0

Charset

utf8

arrayMapFormat

json

Format for map and array inside a column

arrayItemSeparator

3.1.2导入数据后，我们可以看到一个对字段完整性的统计。

我们看到department字段下面的状态条不全是绿的，有一部分是红色（不符合字段类型int），有一部分是白色的（没有数据）。我们可以通过Data Science Studio系统完成对确实字段的数据删除，和对字段类型预测的修改（修改int为text）。

↑

haiku_shirt_sales_1

SummaryExploreChartsStatusSettingsLABACTIONS

Viewing dataset sample

29379 rows, 9 cols

DISPLAY

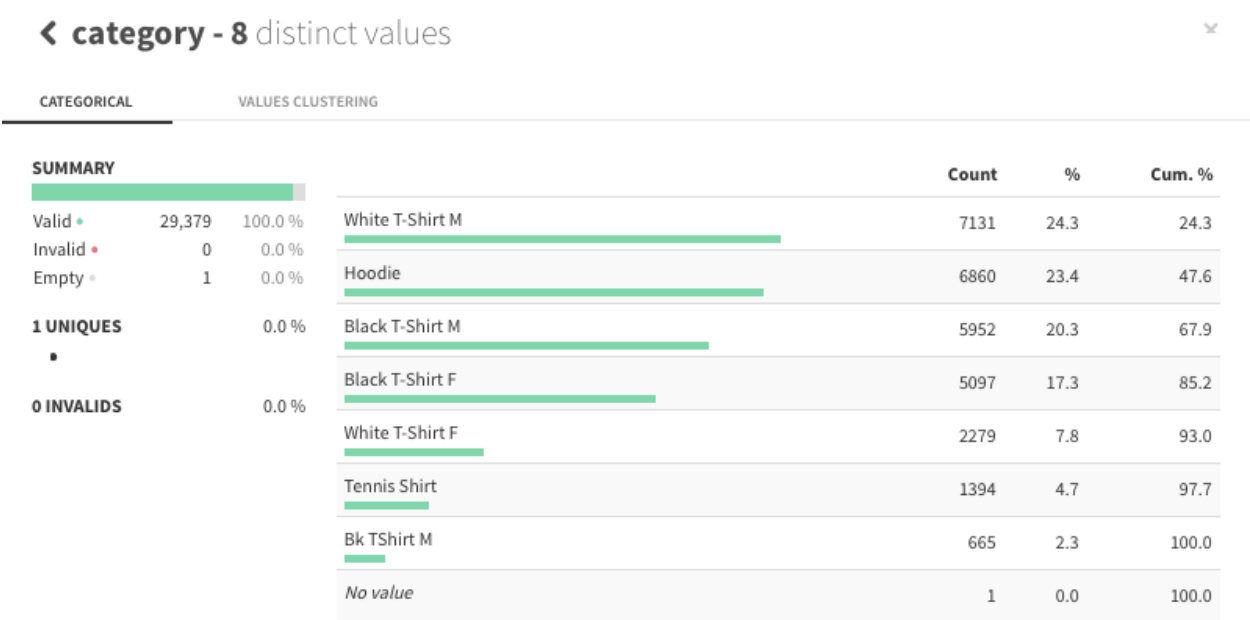
Q

29379 matching rows

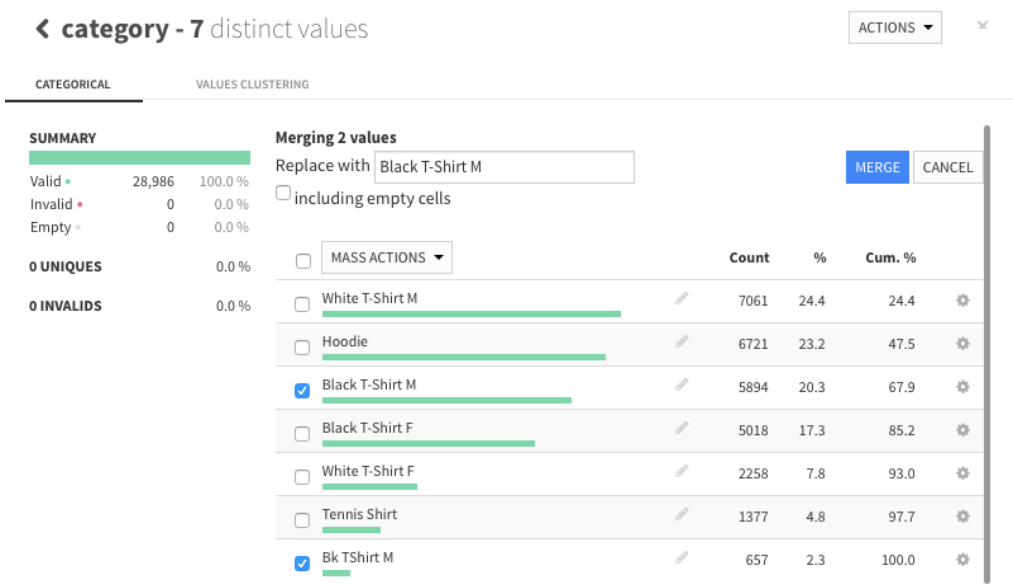
user_id	departement	birth	order_id	order_date	total	nb_tshirts	tshirt_price	category
string Text	string Integer	string Date (unparsed)	string Text	string Date (unparsed)	string Decimal	string Integer	string Decimal	string Text
a4dc1548af		11/9/1970	TSG-1200493	2013-12-07 15:27:00	71.7	3	23.9	White T-Shirt M
e8f6f7853c	91	1/31/1983	TSG-1200500	2014-01-22 08:28:00	19.9	1	19.9	Black T-Shirt M
5802162c20	46	6/5/1958	TSG-1200502	2014-03-22 21:07:00	19.9	1	19.9	Black T-Shirt F
67dc195843	92	12/30/1899	TSG-1200503	2014-09-07 09:03:00	16.7	1	16.7	Black T-Shirt M
7c52e0eab6	78	12/19/1981	TSG-1200505	2014-05-20 15:25:00	16.7	1	16.7	Black T-Shirt F
ef433d39c6	71	1/23/1991	TSG-1200508	2013-12-22 14:09:00	16.7	1	16.7	Black T-Shirt M
17b8b214b0	69	5/24/1982	TSG-1200509	2014-05-22 12:14:00	19.9	1	19.9	White T-Shirt M
1c3dfff940	70	1/5/1971	TSG-1200530	2013-12-05 19:45:00	19.9	1	19.9	Black T-Shirt F
61a7d084f5	75	2/11/1963	TSG-1200534	2013-12-15 15:05:00	33.4	2	16.7	Black T-Shirt M
61a7d084f5	75	2/11/1963	TSG-1200535	2013-12-15 15:05:00	33.4	2	16.7	Black T-Shirt M
ff971288b2	13	5/28/1973	TSG-1200538	2013-11-29 13:19:00	39.8	2	19.9	White T-Shirt M
d72e9b958a	75	12/16/1977	TSG-1200541	2014-02-10 14:22:00	25.0	1	25	Hoodie
87f89d1fe4	53	12/18/1950	TSG-1200545	2014-01-19 10:59:00	200.0	8	25	Hoodie
87f89d1fe4	53	12/18/1950	TSG-1200547	2014-01-19 10:59:00	200.0	8	25	Hoodie
23b5dae249	75	6/7/1977	TSG-1200552	2014-08-20 10:10:00	19.9	1	19.9	Hoodie
82c97f8fbc	13	4/4/1987	TSG-1200555	2013-12-21 20:42:00	75.0	3	25	Hoodie
1208f3603c	52	10/1/1959	TSG-1200560	2013-12-14 12:57:00	23.9	1	23.9	Hoodie
ce7bd68e91	49	2/1/1958	TSG-1200561	2014-03-12 10:03:00	217.1	13	16.7	Black T-Shirt M
50d150a385	93	6/7/1966	TSG-1200564	2014-05-30 21:01:00	16.7	1	16.7	Bk TShirt M
b25f411e02	84	12/30/1899	TSG-1200565	2014-08-14 23:53:00	19.9	1	19.9	Hoodie

3.1.3 字段值分布

我们可以容易地看到各个字段的值得分布



3.1.4 合并相似值



3.1.5 历史记录

对数据集做的改变都会记录下来，可以随意查看或者修改。

Analyze haiku_shirt_sales_1

SummaryScriptChartsModels

DEPLOY SCRIPTACTIONS

Script2 steps

Design Sample29379 rows 9 cols

Script output28986 rows, 9 cols (657 - 393)

DISPLAY

Mass | Search steps...

Remove rows with empty values in departement
393

Replace Bk TShirt M by Black T-Shirt M in category
657

+ ADD A NEW STEP

ADD A GROUP

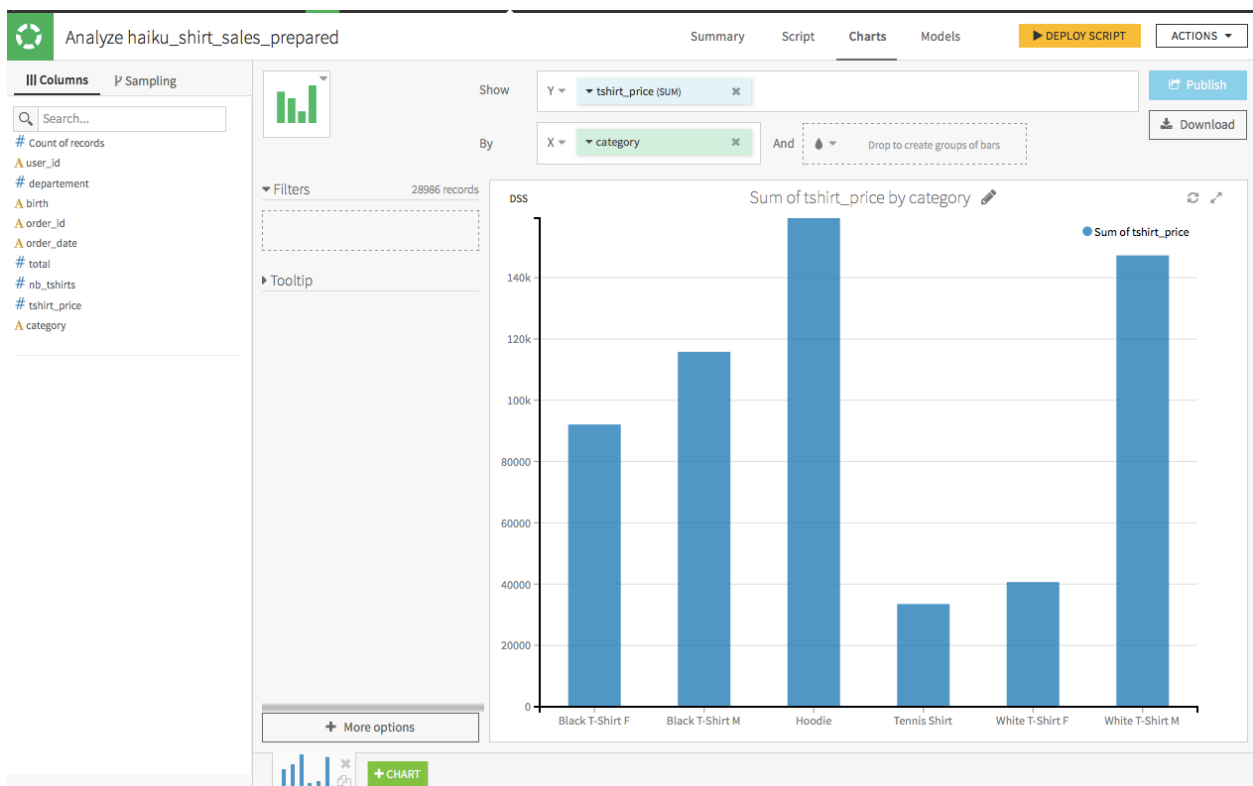
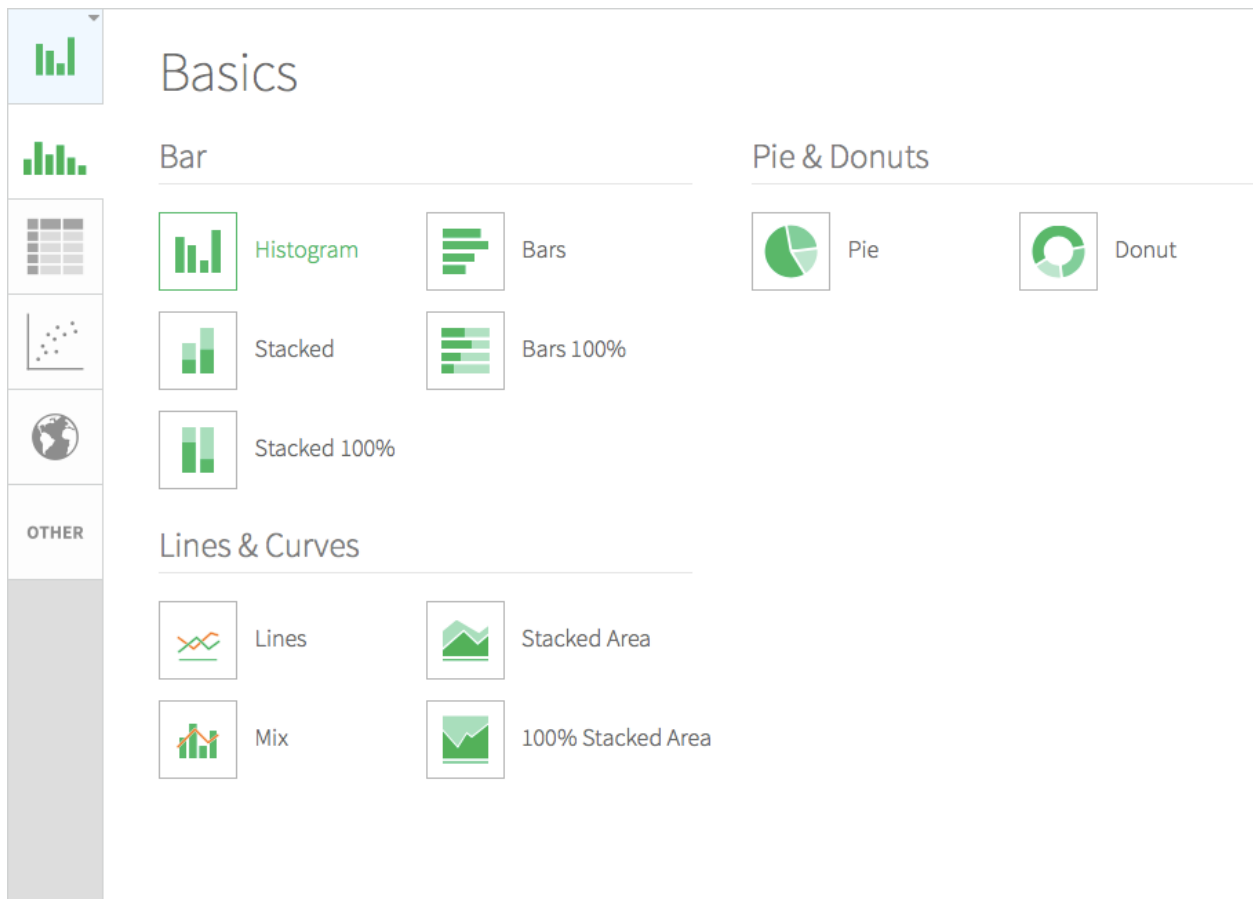
user_iddepartementbirthorder_idorder_datetotalnb_tshirtstshirt_pricecategory

TextTextDate (unparsed)TextDate (unparsed)DecimalIntegerDecimalText

e8f6f7853c	91	1/31/1983	TSG-1200500	2014-01-22 08:28:00	19.9	1	19.9	Black T-Shirt M
5802162c20	46	6/5/1958	TSG-1200502	2014-03-22 21:07:00	19.9	1	19.9	Black T-Shirt F
67dc195843	92	12/30/1899	TSG-1200503	2014-09-07 09:03:00	16.7	1	16.7	Black T-Shirt M
7c52e0eab6	78	12/19/1981	TSG-1200505	2014-05-20 15:25:00	16.7	1	16.7	Black T-Shirt F
ef433d39c6	71	1/23/1991	TSG-1200508	2013-12-22 14:09:00	16.7	1	16.7	Black T-Shirt M
17b8b214b0	69	5/24/1982	TSG-1200509	2014-05-22 12:14:00	19.9	1	19.9	White T-Shirt M
1c3ddff940	70	1/5/1971	TSG-1200530	2013-12-05 19:45:00	19.9	1	19.9	Black T-Shirt F
61a7d084f5	75	2/11/1963	TSG-1200534	2013-12-15 15:05:00	33.4	2	16.7	Black T-Shirt M
61a7d084f5	75	2/11/1963	TSG-1200535	2013-12-15 15:05:00	33.4	2	16.7	Black T-Shirt M
ff971288b2	13	5/28/1973	TSG-1200538	2013-11-29 13:19:00	39.8	2	19.9	White T-Shirt M
d72e9b958a	75	12/16/1977	TSG-1200541	2014-02-10 14:22:00	25.0	1	25	Hoodie
87f89d1fe4	53	12/18/1950	TSG-1200545	2014-01-19 10:59:00	200.0	8	25	Hoodie
87f89d1fe4	53	12/18/1950	TSG-1200547	2014-01-19 10:59:00	200.0	8	25	Hoodie
23b5dae249	75	6/7/1977	TSG-1200552	2014-08-20 10:10:00	19.9	1	19.9	Hoodie
82c97f8fbc	13	4/4/1987	TSG-1200555	2013-12-21 20:42:00	75.0	3	25	Hoodie
1208f3603c	52	10/1/1959	TSG-1200560	2013-12-14 12:57:00	23.9	1	23.9	Hoodie
ce7bd68e91	49	2/1/1958	TSG-1200561	2014-03-12 10:03:00	217.1	13	16.7	Black T-Shirt M
50d150a385	93	6/7/1966	TSG-1200564	2014-05-30 21:01:00	16.7	1	16.7	Black T-Shirt M
b25f411e02	84	12/30/1899	TSG-1200565	2014-08-14 23:53:00	19.9	1	19.9	Hoodie
eda2d9d70a	02	12/9/1979	TSG-1200566	2013-12-07 18:57:00	19.9	1	19.9	White T-Shirt M
c3a94ea2b0	85	6/24/1954	TSG-1200571	2013-12-12 19:21:00	33.4	2	16.7	Black T-Shirt F

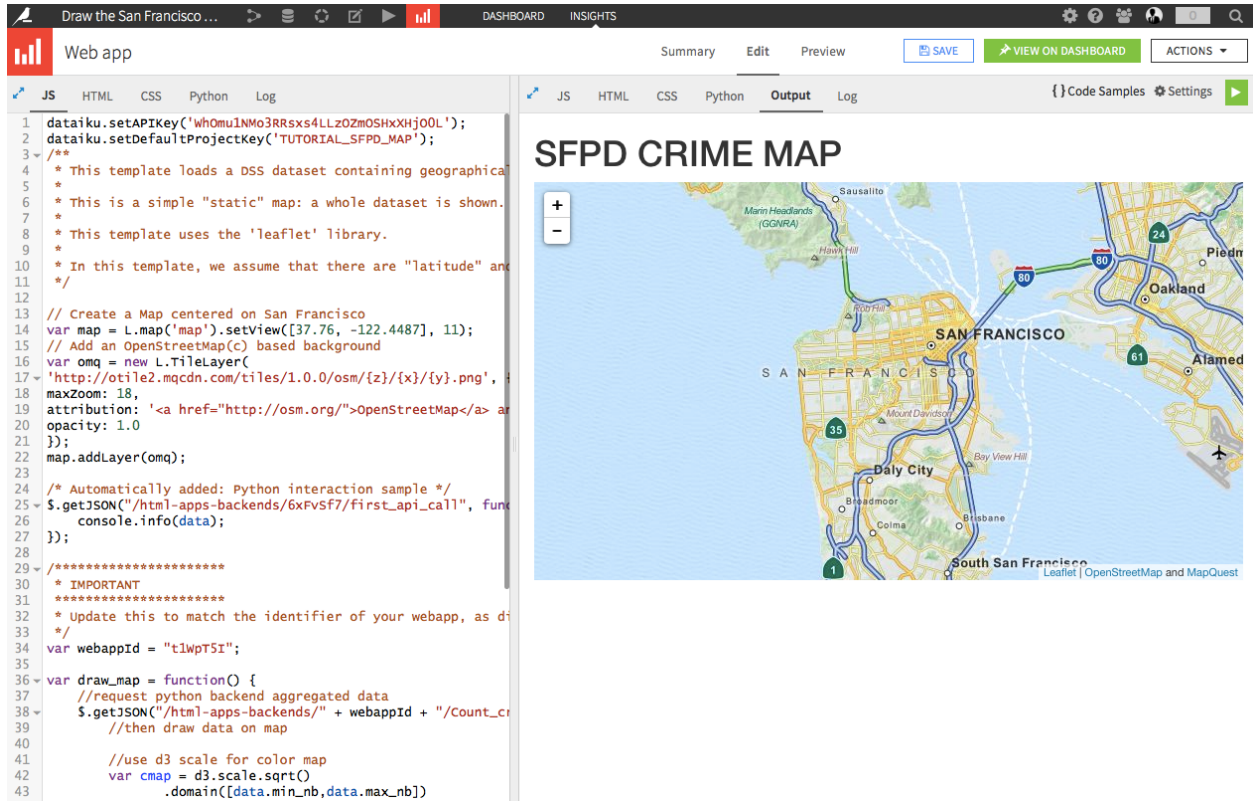
3.2 可视化

可视化展示只提供了一些常用的使用，类似于excel，使用起来还是比较顺手的。



3.3 内容发布

我们可以根据已有的数据集，以及一些网页内容的代码编写，就能轻易发布一个Web应用。



The screenshot displays a web application interface with two main panels. The left panel is a code editor showing JavaScript code for creating a map and fetching data. The right panel is a map titled "SFPD CRIME MAP" showing a map of San Francisco with various landmarks and roads.

Code Editor (Left Panel):

```
1 dataiku.setAPIKey('Wh0mu1NM03RRSxs4LLz0Zm0SHxXHj00L');
2 dataiku.setDefaultProjectKey('TUTORIAL_SFDP_MAP');
3 /**
4  * This template loads a DSS dataset containing geographical
5  *
6  * This is a simple "static" map: a whole dataset is shown.
7  *
8  * This template uses the 'leaflet' library.
9  *
10 * In this template, we assume that there are "latitude" and
11 */
12
13 // Create a Map centered on San Francisco
14 var map = L.map('map').setView([37.76, -122.4487], 11);
15 // Add an OpenStreetMap(c) based background
16 var omq = new L.TileLayer(
17 'http://otile2.mqcdn.com/tiles/1.0.0/osm/{z}/{x}/{y}.png', {
18 maxZoom: 18,
19 attribution: '<a href="http://osm.org/">OpenStreetMap</a> and
20 opacity: 1.0
21 });
22 map.addLayer(omq);
23
24 /* Automatically added: Python interaction sample */
25 $.getJSON("/html-apps-backends/6xFvSf7/first_api_call", function(data) {
26 console.info(data);
27 });
28
29 /*****
30 * IMPORTANT
31 *****/
32 * Update this to match the identifier of your webapp, as d
33 */
34 var webappId = "t1wpt5I";
35
36 var draw_map = function() {
37 //request python backend aggregated data
38 $.getJSON("/html-apps-backends/" + webappId + "/Count_cr
39 //then draw data on map
40
41 //use d3 scale for color map
42 var cmap = d3.scale.sqrt()
43 .domain([data.min_nb, data.max_nb])
```

Map (Right Panel):

The map is titled "SFPD CRIME MAP" and shows a map of San Francisco. The map includes labels for various locations such as Sausalito, Marin Headlands (GGNRA), Hawk Hill, Red Hill, San Francisco, Daly City, South San Francisco, and Oakland. The map also shows major roads like Highway 80 and Highway 61. The map is interactive, with a zoom control in the top left corner.